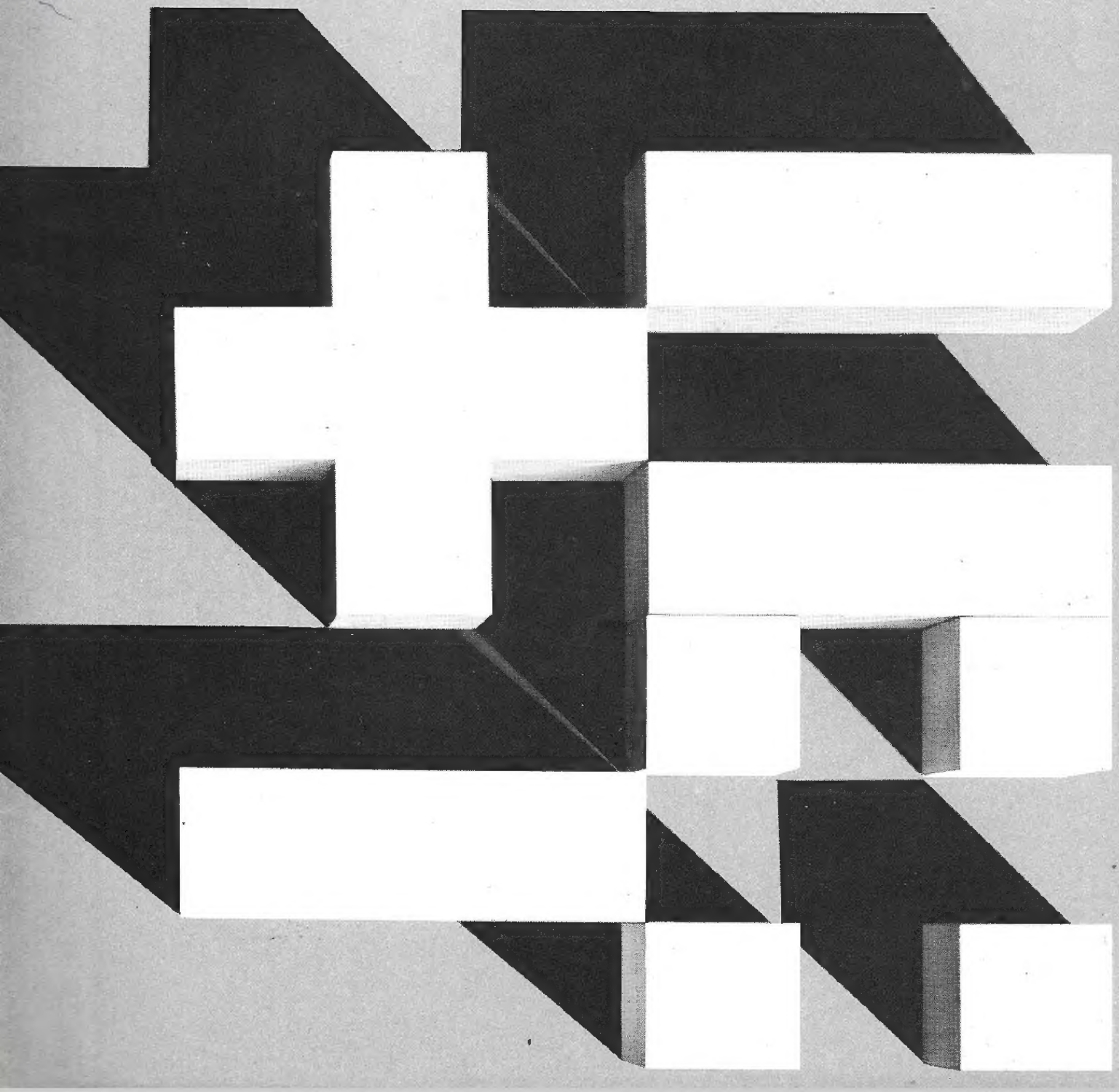
M 100 2

THE OPEN UNIVERSITY

Mathematics Foundation Course Unit 2

Errors and Accuracy



The Open University

Mathematics Foundation Course Unit 2

ERRORS AND ACCURACY

Prepared by the Mathematics Foundation Course Team

Correspondence Text 2

The Open University Press

The Open University Press
Walton Hall Milton Keynes
MK7 6AA

First published 1970. Reprinted 1971, 1972, 1975

Designed by the Media Development Group of The Open University.

Printed in Great Britain by
EYRE AND SPOTTISWOODE LIMITED
AT GROSVENOR PRESS PORTSMOUTH

ISBN 0 335 01001 6

This text forms part of an Open University course. The complete list of units in the course appears at the end of this text.

For general availability of supporting material referred to in this text, please write to the Director of Marketing, The Open University, P.O. Box 81, Walton Hall, Milton Keynes, MK7 6AT.

Further information on Open University courses may be obtained from the Admissions Office, The Open University, P.O. Box 48, Walton Hall, Milton Keynes, MK7 6AB.

1.4

Contents

Objectives

After working through this unit you should be able to:

(i) distinguish between errors of measurement, rounding-off errors and blunders;

(ii) state the meaning of the terms: absolute error
percentage error
error bound
error interval
iterative method
scale factor

(iii) given the absolute error, calculate the relative error and vice versa;

(iv) indicate an error bound in any appropriate standard form, as

$x \pm a$
x expressed to n decimal places
x expressed to n significant figures
$x \in [a, b]$;

(v) given an error in an element in the domain of a function, estimate the error in the image;

(vi) given an error interval in the domain of a function, estimate the corresponding error interval in the codomain;

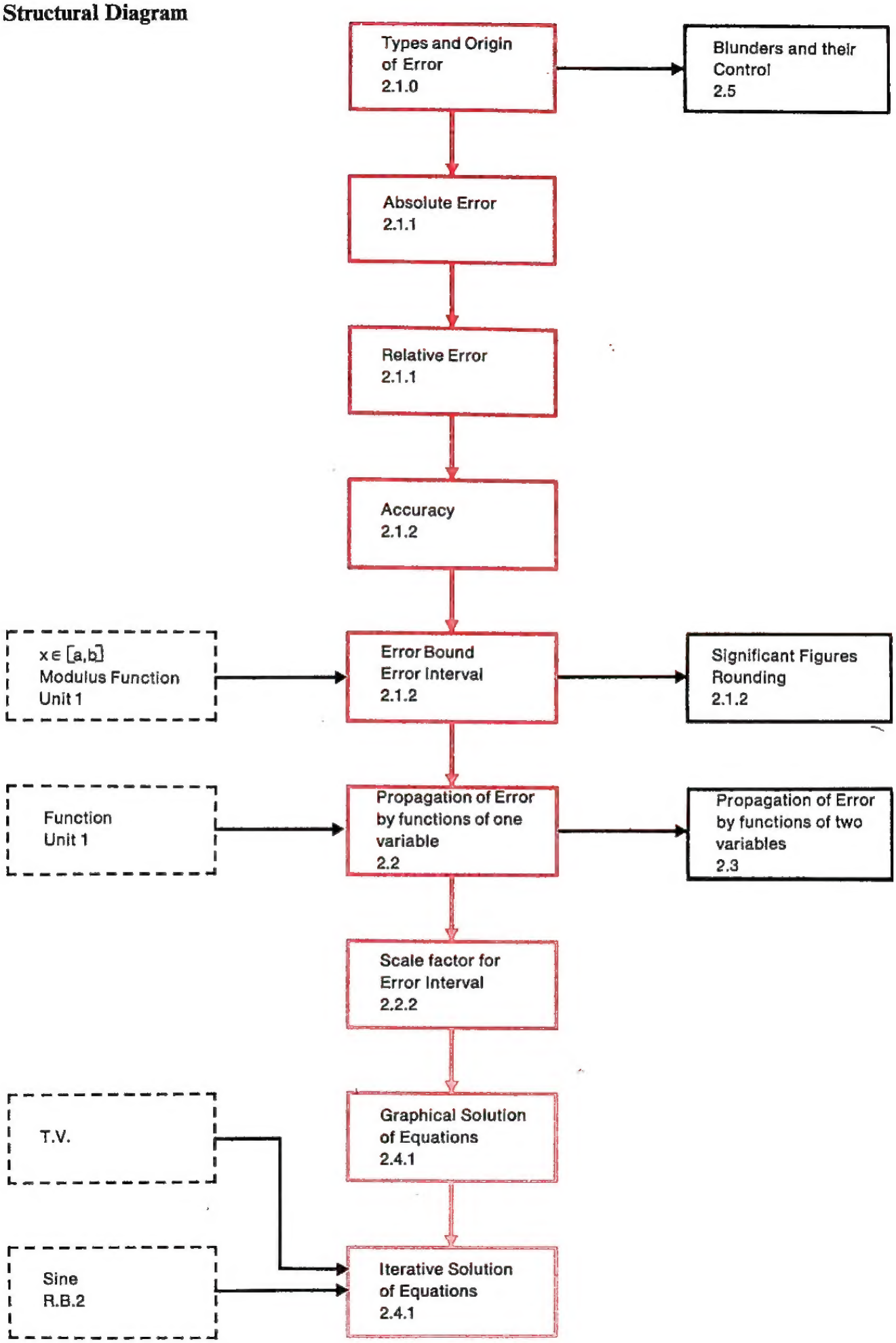
(vii) obtain an approximate solution to $f(x) = 0$, if it exists, using graphical methods, where f is a function;

(viii) rearrange $f(x) = 0$ and derive iterative procedures; test the usefulness of an iterative procedure using the scale factor and obtain a solution, in successful cases, by using the appropriate procedure.

In all cases where a function is mentioned in these objectives, we refer to the fairly simple types of function in the text.

N.B.
Before working through this correspondence text, make sure you have read the general introduction to the mathematics course in the Study Guide, as this explains the philosophy underlying the whole course. You should also be familiar with the section which explains how a text is constructed and the meanings attached to the stars and other symbols in the margin, as this will help you to find your way through the text.

Structural Diagram



Glossary

Terms which are defined in this glossary are printed in CAPITALS.

		Page
ABSOLUTE ERROR	The ABSOLUTE ERROR in a measurement x is the difference, $x - X$, between the measured number x and the exact number X.	3
ABSOLUTE ERROR BOUND	The ABSOLUTE ERROR BOUND in a measurement is the maximum possible value of the MAGNITUDE of the ABSOLUTE ERROR.	6
ERROR INTERVAL	The ERROR INTERVAL is the INTERVAL within which the true value of the quantity must lie.	6
ERROR PROPAGATION	See PROPAGATION OF ERRORS.	
FUNCTION OF ONE (REAL) VARIABLE	A FUNCTION OF ONE (REAL) VARIABLE is a function whose domain is R (or a subset of R) and whose codomain is also R (or a subset of R).	2
FUNCTION OF TWO (REAL) VARIABLES	A FUNCTION OF TWO (REAL) VARIABLES is a function whose domain is a set of pairs of real numbers and whose codomain is R (or a subset of R).	2
INHERENT ERRORS	MEASUREMENT ERRORS and ROUND-OFF ERRORS in the data are together referred to as INHERENT ERRORS (as opposed to blunders).	1
INTERVAL	An INTERVAL is a subset of R consisting of the set of all numbers between, and including, two numbers.	6
ITERATIVE METHOD	An ITERATIVE METHOD for solving a problem is one in which a guess is made at the solution, and a process is repeated over and over again, to try to make the estimate more accurate at each step.	29
MAGNITUDE	The MAGNITUDE of a number x is its image under the modulus function (see *Unit 1, Functions*).	6
MEASUREMENT ERRORS	MEASUREMENT ERRORS are errors arising from the measurement of a physical quantity.	1
PERCENTAGE ERROR	The PERCENTAGE ERROR is the RELATIVE ERROR multiplied by 100.	5
PROPAGATION OF ERRORS	The PROPAGATION OF ERRORS is the way in which errors in the initial data used in a computation affect the final result and any intermediate results.	1
REAL FUNCTIONS	REAL FUNCTIONS are FUNCTIONS OF ONE (REAL) VARIABLE or TWO (REAL) VARIABLES.	2
RELATIVE ERROR	The RELATIVE ERROR in a measurement x (where $x \neq 0$) is the ratio of the ABSOLUTE ERROR to the measured value.	3
RELATIVE ERROR BOUND	The RELATIVE ERROR BOUND in a measurement x (where $x \neq 0$) is the maximum possible value of the MAGNITUDE of the RELATIVE ERROR.	6
ROUND-OFF	To ROUND-OFF a number to n decimal places is to represent the number by the nearest decimal number with n digits after the decimal point.	7
ROUND-OFF ERROR	The ROUND-OFF ERROR of a number is the error introduced by ROUNDING-OFF the decimal representation of the number to a certain number of decimal places.	1

		Page
SCALE FACTOR	The SCALE FACTOR for a FUNCTION OF ONE VARIABLE, propagating an error from the domain to the codomain, is defined as $\frac{\text{estimated error in image of } x}{\text{error in } x} \quad (x \in \text{domain})$	21
SIGNIFICANT FIGURES	A number is expressed to n significant figures if and only if there are n digits from the first non-zero digit in the number to the rounded digit.	8

Notation

The symbols are presented in the order in which they appear in the text.

R	The set of real numbers.	1
e_x	The absolute error in a measurement x.	3
r_x	The relative error in a measurement x, where $x \neq 0$.	3
$a \leqslant b$	a is less than or equal to b (see *Unit 1, Functions*, page 24).	5
$[a, b]$	The interval consisting of all elements of R between, and including, a and b (see also *Unit 1, Functions*, page 24).	5
$x \in A$	The element x belongs to the set A (see *Unit 1, Functions*, page 4).	6
ε_x	The absolute error bound in a measurement x.	6
$\lvert x \rvert$	The image of x under the modulus function (see *Unit 1, Functions*).	6
ρ_x	The relative error bound in a measurement x where $x \neq 0$.	6
$a \pm \varepsilon_a$	An estimate of a, with an absolute error bound ε_a.	6
3.14 9.3×10^7 4×10^{-3}	These are examples of two types of notation that, in this context, are used to express estimates; in the first case, the absolute error bound is 0.005; in the second case, it is $0.05 \times 10^7 = 500\,000$; in the third case, it is $0.5 \times 10^{-3} = 0.0005$.	7 8
$f: x \longmapsto a$	The image of x under the mapping f is a (see *Unit 1, Functions*, page 8).	9
$f \circ g$	The composition of the functions f and g (see *Unit 1, Functions*, page 33).	9
R^+	The set of positive real numbers.	12
P	The set of real functions of the form $x \longmapsto$ *an expression in integer powers of* x.	12
$e_{f(x)}$	The absolute error in $f(x)$, where f is a function propagating an error in x.	12
$a \simeq b$	a is approximately equal to b.	14
$a < b$	a is less than b.	17
$a > b$	a is greater than b.	17

Bibliography

Few, if any, books are written along the lines of this text. Many require a knowledge of calculus and use different definitions. For example, in many of the books the absolute error is defined as

true value − approximate value

rather than our

approximate value − true value

We choose the latter so that later, when some of the ideas are extended to calculus, we shall not be involved in a sign change. In some books, for example B. Noble, *Numerical Methods I* (Oliver and Boyd, 1964), the absolute error is defined as the modulus of our absolute error and similarly for the relative error. For those who already know some calculus, the first two chapters of this book indicate how the subject extends.

Another group of books is fairly closely associated with computer programming and this would tend to be too specialized for your particular use. They also become involved with floating-point arithmetic which you will not meet in this unit. Those interested in the development of the subject in this direction (and who also know calculus) could read Chapters 2 and 5 of D. D. McCracken and W. S. Dorn, *Numerical Methods and Fortran Programming* (John Wiley, 1964).

2.1 THE BASIC CONCEPTS

2.1.0 Introduction

2.1.0 Introduction

This unit is the first of several throughout the course which in part are concerned with the problem of getting numerical answers. In this sense it has a different approach from *Unit 1*. That unit introduced you to some of the precise definitions that go to make up the language of mathematics. Here you will discover how mathematics can deal with the imprecise as well as the precise, and how, having decided what accuracy you wish to achieve in your calculations, you can set about attaining that accuracy.

"Can you guess the weight of the cake?" "How many dried peas in the jar?" Have you ever been asked these questions at a local fête? How accurately can you weigh a cake with your hand? To the nearest kilogram? Can you estimate the number of dried peas to the nearest 100? Embodied in these simple examples is the idea that many of the numbers that occur in our life are approximations. At what time did you leave the house this morning? How many cars did you see on your way to work? It is doubtful whether you could answer either of these questions precisely. The law recognizes that inaccuracies are inevitable. If you obtain 5 gallons of petrol from a petrol pump which has been in use for some time, legally the quantity you get may be between $4\frac{63}{64}$ and $5\frac{1}{16}$ gallons. Even the extremely precise atomic clocks have a possible inaccuracy of 5 seconds in 700 years.

The various quantities quoted above fall into two types. One type comprises the quantities, like the weight of the cake, that we can never find precisely; no matter how fine a balance we use there is always a small possible error in the measurement of the weight. The other type comprises the quantities, like the number of peas in the jar, that we can, in principle, find precisely by counting. In this text we are concerned mainly with quantities of the first type. All these are examples of measurement errors arising because some measurement of a physical quantity is not perfectly accurate. Measurements are not, however, the only source of inaccuracies: try writing π or $\frac{1}{3}$ as an exact decimal. However many decimal places you write down there must be some error in your representation. Such an error, arising from the fact that the number is not given exactly by the decimal representation used, is called a round-off error. Errors of measurement and round-off errors in data have a similar effect when the data are used in a calculation. We refer to the two types of error collectively as inherent errors.

Definition 1

Definition 2

Definition 3

Let us take the above example of the dried peas in the jar a little further. Suppose you are not allowed to count them but would like to improve on a simple guess. So you argue something like this: on average, a dried pea looks as if it is $\frac{1}{2}$ cm in diameter. The jar has a square base with sides about 8 cm long, and about 10 cm high. Thus you deduce that the number is somewhere in the region of

$$\frac{8}{\frac{1}{2}} \times \frac{8}{\frac{1}{2}} \times \frac{10}{\frac{1}{2}} = 5120$$

Let us analyse briefly what you have done. You have used inaccurate data in an exact computation and derived, as you know, an inaccurate result. The question arises: "How inaccurate is the result?" and this takes us into the topic of the propagation of errors. By this we mean the way in which errors in the initial data used in a computation affect the final result and any intermediate results.

Definition 4

This topic is introduced in this unit by investigating firstly the propagation of errors by functions with domain and codomain R, the set of real numbers, and then later in the unit by functions whose domain is a set

of pairs of real numbers (or triples, etc.) and codomain R, as in the example above, where there may be errors in our estimates of the height of jar, base of jar, diameter of pea.*

We will then see how familiarity with the ideas of error propagation enables us to solve a particular class of purely mathematical problem — the numerical solution of equations — by improving guesses until we attain a desired accuracy.

Finally we will investigate some of the ways in which we can control a type of error to which we are all rather prone — the blunder.

Exercise 1

Exercise 1
(2 minutes)

How many dried peas would you estimate to be in the jar in the previous example if you assumed the diameter of the pea to be 0.4 cm? Are you surprised at the different result you obtain? ■

* Functions whose domain and codomain are R or a subset of R are called functions of one (real) variable; functions whose domain is a set of pairs of real numbers and codomain is R (or a subset of R) were introduced in the previous unit and are called functions of two (real) variables.

Definition 5
* *

When it is clear from the context which of these two types of functions we are referring to, we sometimes use an alternative shortened form, and call them real functions.

Definition 6
* *

Definition 7
* *

2.1.1 What is Error?

2.1.1 Discussion

It may seem surprising that a mathematical treatment of errors can exist. One naturally thinks of mathematics as an exact discipline, in which errors can arise only through mistakes or imperfections which should not be tolerated in mathematical work. This is a misconception, however; provided we can define precisely what we mean by an "error" and attach a numerical value to it, we can apply mathematical reasoning to the errors just as we do with any of the other objects to which we apply mathematics.

To define "error" mathematically, let us suppose that we are using one number, which we denote by x, as an approximation to another, which we denote by X. For example, the "exact" number X might be $\frac{1}{3}$ and x the approximation 0.33; or X could be π and x the approximation $\frac{22}{7}$; or X could be the actual number of peas in a jar and x your guess at this number; or X could be the precise amount of petrol you received and x the amount as measured by the meter on the petrol pumps. In each case, the numbers X and x are likely to be different, and we define their difference $x - X$ as the absolute error in x and denote it by e_x:

Definition 1

Notation 1

$$e_x = x - X$$

The word "absolute" is to distinguish this measure of error from another one, called the "relative error", which we shall meet presently. (In general, we use just "error" when it is clear from the context which we mean.) Note that e_x can be either positive or negative, according as the approximation x is larger or smaller than the exact value X. As an example, if we imagine we have a worn tape measure with the first centimetre missing, then if the true length being measured were 14 cm, we would, assuming for a moment that the rest of the tape measured exactly, record a value of 15 cm. This we would call the approximate value in this instance. Thus

$$15\text{ cm} \quad - \quad 14\text{ cm} \quad = \quad 1\text{ cm}$$

(approximate value − true value = absolute error)

and the absolute error would be 1 cm.

In other words, to correct the values given by the tape, we must always make a correction (the negative of the error) of −1 cm, i.e. we *subtract* the error.

For another example, if you know that your speedometer consistently records 5 mile/h too high, a recorded (approximate) value of 37 mile/h would correspond to a true value of 32 mile/h with an absolute error of 5 mile/h and a correction needed of −5 mile/h, i.e.

$$37\text{ mile/h} \quad - \quad 32\text{ mile/h} \quad = \quad 5\text{ mile/h}$$

(approximate value − true value = absolute error)

In the example of the tape measure above, we had an error of 1 cm in 15 cm. It is possible to measure a distance of 1 km to an accuracy of 1 cm, i.e. the possible absolute error is again 1 cm in a recorded value of 1 km. Clearly this second measurement is, in a sense, more accurate than the first, although the absolute error is the same in each case. To allow for this type of distinction we use the relative error in x defined as $\frac{e_x}{x}$ and written as r_x:

Definition 2

Notation 2

$$r_x = \frac{e_x}{x}$$

(Note that we compare e_x with x, the approximate value, because in general we know this value and not the exact value X.)

Solution 2.1.0.1

Solution 1

$$\frac{8}{0.4} \times \frac{8}{0.4} \times \frac{10}{0.4} = 10\,000$$

It is quite surprising that the 20% change in the supposed diameter of the pea nearly doubles the estimate of the number. ■

Thus in the above two examples we have approximate value $x = 15$ cm, absolute error $e_x = 1$ cm, giving

$$\text{relative error } r_x = \frac{1}{15}$$

and approximate value $x = 10^5$ cm, absolute error $e_x = 1$ cm, giving

$$\text{relative error } r_x = \frac{1}{10^5} = 10^{-5}$$

showing how much smaller the relative error is in the second case. Multiplied by 100, the relative error is the percentage error you have probably met before. Often knowledge of the relative, or percentage, error is more useful than knowledge of the absolute error, since it gives a measure of the error in relation to the size of the number being considered. This is not always the case however; for example, the absolute error in the diameter of an axle is clearly the more important when we are fitting it into a ball-race. And if the approximate value of the number is zero (as when measuring the oxygen content of polluted river water!), the definition of relative error loses its meaning.

Definition 3

2.1.2 What is Accuracy?

2.1.2 Discussion

The naive answer to the question: "What is accuracy?" is that it is simply the absence of error — i.e. that a small error corresponds to a high accuracy and vice-versa. There is a lot of truth in this, but it is not the whole story. Suppose you bought a nominal 5 gallons at each of two apparently identical petrol pumps designed to comply with the legal requirement mentioned earlier, i.e. that the true amount must lie between $4\frac{63}{64}$ and $5\frac{1}{16}$ gallons, and that at one pump you happened by chance to get $5\frac{1}{160}$ gallons, and at the other you got the legal maximum, $5\frac{1}{16}$ gallons. At the first pump the error was $\frac{1}{160}$ gallon and at the second it was $\frac{1}{16}$ gallon, but would it be reasonable to say that one pump was ten times as accurate as the other on this account alone? On another day, the position might be reversed, purely by chance. We would like to frame our definition of accuracy so as to be independent of such caprices.

We can do this by making the definition of accuracy depend on the pump itself and not on the amount it delivers on any particular occasion. That is, the accuracy is defined by specifying, not the error on any particular occasion, but bounds between which the error must lie. In the case of the petrol pump, the accuracy in measuring 5 gallons must satisfy the legal requirement that x, the amount of petrol delivered, must lie between $4\frac{63}{64}$ and $5\frac{1}{16}$. By our definition of the absolute error, $e_x = x - X$, this condition requires that e_x lie between $-\frac{1}{64}$ and $+\frac{1}{16}$. In symbols, this is

$$-\tfrac{1}{64} \leqslant e_x \leqslant \tfrac{1}{16}$$

It can also be written

$$e_x \in [-\tfrac{1}{64}, \tfrac{1}{16}]$$

where $[-\frac{1}{64}, \frac{1}{16}]$ is the set consisting of all numbers from $-\frac{1}{64}$ to $\frac{1}{16}$ inclusive. Such a set is called an interval. The interval $[4\frac{63}{64}, 5\frac{1}{16}]$ is called the error interval.

Notation 1
Definition 1
Definition 2

This provides the answer to our question: "What is accuracy?" The accuracy of an approximate number is specified by giving an interval within which the error in the number must lie. The reason why we specify the accuracy in this way, rather than by giving the error itself, is that we do not normally know the error — if we did, we could just subtract it from the approximate number and recover the exact number.

It frequently happens (though not in the petrol pump example) that the interval used to specify the accuracy is symmetrical about zero so that the condition on the error in a number x has the form

$$e_x \in [-\varepsilon_x, \varepsilon_x]$$

where ε_x is some positive number.

This condition can also be written

$$|e_x| \leqslant \varepsilon_x$$

Notation 2

where the e_x between vertical bars denotes the *magnitude* of the number e_x (its image under the modulus function defined in *Unit 1, Functions*). When the accuracy is specified by a symmetrical interval like this, we call the number ε_x the absolute error bound of x. An alternative way of specifying the accuracy of an approximate number x is to use the relative error bound defined by

Definition 3
Definition 4

$$\rho_x = \frac{\varepsilon_x}{|x|}$$

Notation 3

so that the relative error r_x satisfies

$$|r_x| \leqslant \rho_x$$

Exercise 1

Exercise 1
(2 minutes)

We record a measurement of 2.5 kg and assume that there is a maximum error in the instrument of 0.05 kg, that is, the true value is in the interval [2.45, 2.55] kg.

What is

(i) the absolute error bound?
(ii) the relative error bound? ■

A common notation for specifying absolute error bounds is to write, for example,

$$\pi = 3.14 \pm 0.005$$

Notation 4

to indicate that 3.14 is an approximate value for the exact number π and that the absolute error bound is 0.005.

Another method of specifying error bounds depends on the convention for rounding off decimals. You probably know already how to round off a decimal to fewer places. For example, the number π to 10 decimal places is

$$3.1415926536\ldots$$

To save writing and arithmetical labour we very rarely work with this value, but use the best approximation obtainable with, say, 2 or 4 decimal places. The two-place approximation is

$$3.14$$

since this has an error

$$3.14 - 3.1415926536\ldots = -0.0015926536\ldots$$

whereas any other two-place approximation, say 3.13 or 3.15, would have a larger error. In this case the two-place approximation is identical with the first 3 digits of the exact (non-terminating) decimal for π. With four places, on the other hand, the best approximation is

$$3.1416$$

since the error

$$3.1416 - 3.1415926536\ldots = 0.0000073464\ldots$$

is smaller than for (say) 3.1415 or 3.1417. This procedure of representing a number by the closest decimal with some given number, say n, of digits after the decimal point, is called rounding-off the number to n decimal places.

Definition 5

Exercise 2

Exercise 2
(2 minutes)

Round off the following numbers:

(i) $\frac{1}{3}$ to 3 decimal places,
(ii) π to 6 decimal places,
(iii) 0.9999 to 3 decimal places. ■

If an exact number, X, is approximated by its round-off form with n decimal places, x_n, the absolute error bound is

$$\overbrace{0.00\ldots05}^{n \text{ zeros}}$$

since

$$X \in [x_n - \overbrace{0.00\ldots05}^{n \text{ zeros}},\ x_n + \overbrace{0.00\ldots05}^{n \text{ zeros}}]$$

This shows that any rounded-off decimal implies an error bound, and so we can use rounded-off decimals to specify the accuracy of an approximation without giving the error bound explicitly. Thus we write

$$\pi = 3.14$$

Notation 5

or

$$\pi = 3.14 \text{ to two decimal places}$$

to mean that the approximation 3.14 has the error bound characterizing two-place accuracy, i.e. that

$$\pi = 3.14 \pm 0.005$$

There is one convention that should be mentioned here. When rounding a number to one less figure we increase the previous digit by 1 if the last digit is 6, 7, 8 or 9. If the last digit is 0, 1, 2, 3 or 4, we leave the previous

Convention

(*continued on page 8*)

Solution 1

Solution 1

The approximate value x is 2.5 kg.

(i) Maximum magnitude of e_x is 0.05 kg = absolute error bound.

(ii) Maximum magnitude of the relative error is

$$\frac{0.05\text{ kg}}{2.5\text{ kg}} = 0.02 = \text{relative error bound.}$$ ■

Solution 2

Solution 2

(i) 0.333, (ii) 3.141593, (iii) 1.000. ■

(*continued from page 7*)

digit untouched. If the last digit is a 5, the convention is that we look at the previous digit: if it is *even* we leave it unchanged, if it is *odd* we increase it by 1. This is to avert any bias in always rounding to the larger number.

If we round a number to two less figures, we look at the last two digits as a pair, and round according to the rule:

> if the last two digits are less than 50, leave previous digit untouched,
> if the last two digits are greater than 50, increase previous digit by 1.
>
> E.g. 5.371 becomes 5.4
> and 5.329 becomes 5.3.

What would you suppose happens if the last two digits are 50?

Discussion

Sometimes the term significant figures is used instead of "the number of decimal places". For example, we say that the number

Definition 6

12.04

has four significant figures — two in front of the decimal point and two after. The number

0.0001204

also has four significant figures, the last four. (The first three zeros only serve to distinguish the number from 0.1204, for example, and are not said to be significant.) In the statement

"The sun is 93 000 000 miles from the earth"

only the first two figures are significant: the statement means that the distance of the sun from the earth is closer to 93 000 000 than to 94 000 000 or 92 000 000, not that it is closer to 93 000 000 than to 93 000 001 or 92 999 999. To avoid ambiguity in these cases it is convenient to write such numbers in the form

$$9.3 \times 10^7$$

Notation 6

which makes it clear that there are just 2 significant figures, so that the absolute error bound is 0.05×10^7 or 500 000.

(*continued on page 11*)

Exercise 3

Exercise 3
(2 minutes)

(i) Indicate which of the given answers is correct. The statement "$a = 12.43 \pm 0.01$" means:
   (a) a varies in steps of 0.01 between 12.42 and 12.44.
   (b) a has some value between and including 12.42 and 12.44.
   (c) a is equal to 12.42 or 12.44.

(ii) To what number of significant figures are the following numbers given?
   (a) 28.237, (b) 0.0474, (c) 125.0×10^3. ■

Exercise 4

Exercise 4
(1 minute)

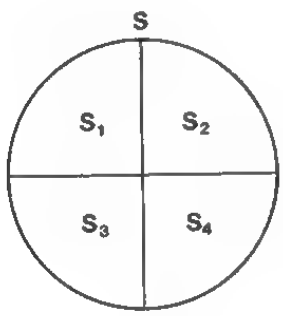
(5 minutes)

(i) What is the absolute error bound if you measure the width of a pane of glass to the nearest centimetre?

(ii) Determine ε_x and ρ_x given
   (a) $x = 25$ min with a maximum possible error of 6 sec in the clock.
   (b) $x = 40$ min with a maximum percentage error of 5%.
   (c) $x = 0.05$ after rounding to 2 places of decimals. ■

Exercise 5

Exercise 5
(5 minutes)



Suppose we consider the set, S, of all numbers of four significant figures or less, split up into four subsets.

S_1 contains all numbers with one significant figure.
S_2 contains all numbers with two significant figures.
S_3 contains all numbers with three significant figures.
S_4 contains all numbers with four significant figures.

We define the functions

$f_1 : x \longmapsto$ (x rounded to one significant figure less)
($x \in S_2$ or S_3 or S_4),

$f_2 : x \longmapsto$ (x rounded to two significant figures less)
($x \in S_3$ or S_4).

(i) What is the image of the domain of f_1?
(ii) What is the image of the domain of f_2?
(iii) How many numbers map to the number 4 under f_1?
(iv) How many numbers map to the number 4 under f_2?
(v) Is the following statement true or false?

$$f_1 \circ f_1(x) = f_2(x) \quad (x \in S_3 \text{ or } S_4)$$ ■

Solution 3 **Solution 3**

(i) (b) In the context of error the convention for this notation is that it means *some* value in the interval $[12.43 - 0.01, 12.43 + 0.01]$. For example, measure the page in front of you. How accurate are your measurements? With a normal ruler you can measure the paper within 0.1 cm, possibly more precisely if you try hard. Consequently, a measurement recorded as 21.1 cm could represent any length between 21.0 cm and 21.2 cm, and is written as 21.1 ± 0.1 cm.

(c) This answer would be correct in a different context; that is, when solving $(a - 12.43)^2 = 0.0001$ we could write the solution as 12.43 ± 0.01.

(ii) (a) 5, (b) 3, (c) 4. ■

Solution 4 **Solution 4**

(i) 0.5 cm.

(ii) (a) $\varepsilon_x = 6$ sec, $\rho_x = 4 \times 10^{-3}$,
(b) $\varepsilon_x = 2$ min, $\rho_x = 5 \times 10^{-2}$,
(c) $\varepsilon_x = 5 \times 10^{-3}$, $\rho_x = 10^{-1}$. ■

Solution 5 **Solution 5**

(i) S_1, S_2 and S_3.

(ii) S_1 and S_2.

(iii) 11, these are 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5. (Remember the convention on page 7.)

(iv) 101, these are 3.50, 3.51, 3.52, . . . , 4.47, 4.48, 4.49, 4.50.

(v) False. For example, if $x = 3.47$ we have

$$f_1(3.47) = 3.5, \quad f_1(3.5) = 4,$$

but

$$f_2(3.47) = 3$$ ■

(*continued from page 8*)

Discussion

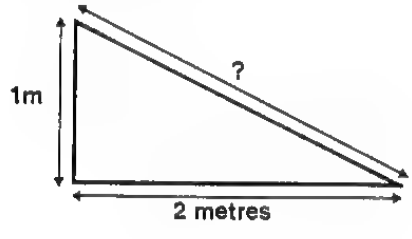
Accuracy is always an important objective, but we must take care not to claim to have attained more of it than in fact we have. Very often, even where one can theoretically attain a high accuracy, it may not be worth while. For example, if your temperature were recorded as 39.962 °C rather than 40 °C the extra digits would convey no more information to you or the doctor. You would just be very sick. In any case shortly afterwards the temperature could well have changed considerably from the former value whilst remaining approximately 40 °C.

Care must always be taken in any calculation (both to ensure the credibility of the result and to save work) to quote the result only to the accuracy implied by the data and the calculation process. You may not be able to do this precisely now, but at least you should be able to recognize when the result is clearly overstated. This is frequently referred to in the sciences as recognizing the *order of magnitude* of the errors involved. An instance of striving for accuracy which is unattainable is illustrated by the following exercise.

Exercise 6

Exercise 6
(2 minutes)

In the right-angled triangle shown we measure the height as 1 metre and the base as 2 metres, both measurements being accurate to the nearest centimetre. By the theorem of Pythagoras the length of the hypotenuse can be calculated as 2.23607 metres. Is this a sensible deduction?



■

Solution 6

Solution 6

NO. The answer quoted in the question is $\sqrt{5}$ correct to five places of decimals, but this is meaningless in the context of the question, since this implies five-figure accuracy in the hypotenuse, whereas the original data had an accuracy only to the nearest centimetre. A reasonable answer would be 2.24 metres. ■

2.2 HOW FUNCTIONS OF ONE VARIABLE PROPAGATE ERRORS

2.2

2.2.0 Introduction

2.2.0

Introduction

In mathematics we are concerned, not only with numerical data, but also with calculations that may be performed on the numbers forming the data. If there are errors in the data, they will affect the result of the calculation, and so the accuracy of the result depends on the accuracy of the data. The mathematical theory of errors makes it possible to express the accuracy of the result of a given calculation in terms of the accuracy in the data. In this section you will learn how to do this in the simple case where the calculation in question is the evaluation of images of a function such as

$$f: x \longmapsto 3x^2 - 2x + 1 \qquad (x \in R)$$

or

$$g: x \longmapsto 5x^4 - 6x + \frac{1}{x} - \frac{7}{x^3} \quad (x \in R, x \neq 0)$$

or

$$h: x \longmapsto \frac{1}{1 + x} \qquad (x \in R^+)$$

i.e. those with domain and codomain R, or a subset of R, and in which the formula specifying the rule contains only integer powers of x in the numerator and/or the denominator. We will call the set of such functions P.

Notation 1

In Section 2.1 we listed and examined some of the errors that can occur in numbers which we may have to use in subsequent calculations. These subsequent calculations often take the form of evaluating the images of such numbers under a function from P. The question we wish to answer is: What happens to these errors when we evaluate these images? Given an error e_x in the original number, what is the error in the image? Suppose for the moment that we know the true value X of the original number and that the approximate value is

$$x = X + e_x$$

Diagrammatically we have

$$x(= X + e_x) \qquad f(x)(= f(X) + ?)$$

The exact value of the image is $f(X)$, and the approximate value we obtain if we use x instead is $f(x)$, so that the error in the image is

$$e_{f(x)} = f(x) - f(X)$$

Equation (1)

For example, if

$$f: x \longmapsto x^2 \qquad (x \in R)$$

with $X = 1, e_x = 0.1$, then we have

$$1.1(= 1 + 0.1) \rightarrow \boxed{\text{Square it}} \rightarrow 1.21(= 1 + ?)$$

and the error in the image is 0.21.

But in fact it would be very tedious and clumsy to have to calculate $e_{f(x)}$ in many cases. Very often the errors are small compared with x (i.e. the relative error is very small), so that if, for instance, the square of the error occurs in $e_{f(x)}$ it will be smaller still. We shall see in the following examples that we can often usefully simplify the formula for the error in the image and obtain a satisfactory estimate much more quickly than by using Equation (1).

2.2.1 Basic Operations of Multiplication, Division, Addition and Subtraction

2.2.1
Main Text

Multiplication by an Exact Number

If

$$f: x \longmapsto 5x \quad (x \in R)$$

then we can represent the action of a function f by the diagram

$$x = (X + e_x) \rightarrow \boxed{x \longmapsto 5x} \rightarrow 5X + 5e_x$$

and represent the corresponding absolute errors in the domain and codomain schematically by

$$e_x \longrightarrow 5e_x = e_{5x}$$

> To find the absolute error in the image we simply multiply the absolute error in the original number by the appropriate factor.

Rule 1

Other Products

Discussion

Consider the function "square it".

$$f: x \longmapsto x^2 \quad (x \in R)$$

Then

$$x = (X + e_x) \rightarrow \boxed{x \longmapsto x^2} \rightarrow (X^2 + 2Xe_x + e_x^2)$$

and

$$e_x \longrightarrow 2Xe_x + e_x^2 = 2xe_x - e_x^2 = e_{x^2}$$

The second expression on the right-hand side is found by substituting

$$X = x - e_x$$

in the second term of the first expression. Again, with the particular numerical values $X = 1$, $e_x = 0.1$, we find

$$(1 + 0.1) \longrightarrow [\; x \longmapsto x^2 \;] \longrightarrow 1 + 2 \times 1.1 \times 0.1 - 0.01$$
$$= 1 + 0.22 - 0.01$$
$$= 1 + 0.21$$

Note the relative sizes of the terms on the right. The number 0.01 is small compared even with the total error 0.21. This shows us where we can gain in simplicity with the marginal loss of some accuracy. Provided e_x *is small compared with* x, e_x^2 will always be *much smaller* than $2xe_x$ and we can safely ignore it. Thus we can usefully say that the error in x^2, e_{x^2}, is about $2xe_x$, i.e.

$$x \longrightarrow [\; x \longmapsto x^2 \;] \longrightarrow x^2$$

$$e_x \longrightarrow \text{estimated } e_{x^2} = 2xe_x$$

If we look at the behaviour of the relative error for the same function, we get the very simple rule

$$r_{x^2} \simeq \frac{2xe_x}{x^2} = 2r_x \quad (x \neq 0)$$

(The symbol $\simeq$ means "approximately equal to")

> Squaring an approximate number roughly doubles its relative error.

Rule 2

Exercise 1

Exercise 1
(5 minutes)

(i) Find a useful estimate of the absolute error in x^3, if the absolute error in x, e_x, is small.
(ii) Express r_{x^3} approximately in terms of r_x, assuming $x \neq 0$.
(iii) What would you think would be useful approximations to e_{x^n} and r_{x^n}?

■

From the result of the last exercise we can conjecture an important principle, i.e. that in multiplication we can *add* relative errors to obtain an *estimate* of the relative error in the product. For example, we can express x^5 as the product of x^3 and x^2, and by (iii) of the last exercise, we have

$$r_{x^5} \simeq 5r_x, \qquad r_{x^3} \simeq 3r_x \quad \text{and} \quad r_{x^2} \simeq 2r_x$$

so that the conjecture, which gives $r_{x^5} \simeq r_{x^3} + r_{x^2}$ is verified in this case. This is an important point to remember for future use.

> For an estimate of the relative error in multiplication, add the relative errors.

Rule 3

Division

Consider

$$f : x \longmapsto \frac{1}{x} \qquad (x \in R, x \neq 0)$$

Then

$$(X + e_x) \longrightarrow \boxed{x \longmapsto \frac{1}{x}} \longrightarrow \frac{1}{X + e_x} = \frac{1}{X} + ?$$

By some algebraic manipulation (you need not derive this, just check it if you wish),

$$e_{1/x} = \frac{1}{X + e_x} - \frac{1}{X} = \frac{-e_x}{(X + e_x)X} = \frac{-e_x}{x(x - e_x)}$$

We ignore the e_x in the denominator by comparison with the x next to it, and get

$$e_{1/x} \simeq -\frac{e_x}{x^2}$$

and

$$\boxed{r_{1/x} = \frac{e_{1/x}}{1/x} \simeq -\frac{e_x}{x} = -r_x}$$

Rule 4

Example 1

Example 1

By writing

$$\frac{1}{x^2} = \left(\frac{1}{x}\right)^2$$

estimate r_{1/x^2} and e_{1/x^2}.

By using Rule 3 for error estimates in multiplication we find

$$r_{1/x^2} \simeq 2r_{1/x}$$

and hence

$$r_{1/x^2} \simeq -2r_x$$

by Rule 4.

Therefore, we have

$$e_{1/x^2} = \frac{1}{x^2}(r_{1/x^2}) \simeq -\frac{2r_x}{x^2} = -\frac{2e_x}{x^3}$$

Note that if we want the absolute error estimate of a product, it is simpler to find the relative error estimate first by the simple Rule 3 we have developed on page 14. ■

Main Text

Addition and Subtraction

It is fairly clear that for these operations we simply add (or subtract) the appropriate absolute error estimates. Thus, for example, consider the function

$$f : x \longmapsto x^3 + x^2 + x \quad (x \in R)$$

The absolute error in the image is

$$e_{x^3} + e_{x^2} + e_x \simeq (3x^2 + 2x + 1)e_x$$

Two points emerge from this:

(i) The absolute error in a sum is equal to the sum of the absolute errors in its terms.

Rule 5

(*continued on page 16*)

Solution 1

Solution 1

(i) $(X + e_x) \longrightarrow \boxed{x \longmapsto x^3} \longrightarrow (X^3 + 3X^2e_x + 3Xe_x^2 + e_x^3)$

$$e_x \longrightarrow 3X^2e_x + 3Xe_x^2 + e_x^3$$

$$= 3(x - e_x)^2e_x + 3(x - e_x)e_x^2 + e_x^3$$

$$= 3x^2e_x \boxed{\begin{array}{c}\text{small}\\ -3xe_x^2 + e_x^3\end{array}}$$

Since e_x is small compared with x, a useful estimate for e_{x^3} is

$$e_{x^3} \simeq 3x^2e_x$$

(ii) $r_{x^3} = \dfrac{e_{x^3}}{x^3} \simeq \dfrac{3x^2e_x}{x^3} = \dfrac{3e_x}{x} = 3r_x$

(iii) Generalizing from x^2 and x^3 suggests

$$e_{x^n} \simeq nx^{n-1}e_x, \qquad r_{x^n} \simeq nr_x \qquad (x \neq 0)$$

This is an important result, which can be justified using the Binomial Theorem. ■

(*continued from page 15*)

(ii) For addition and subtraction, even if we wished to find the relative error, it is simpler to find the absolute error first.

Rule 6 * * *

Thus, in the above, the estimated relative error would be

$$\frac{3x^2 + 2x + 1}{x^3 + x^2 + x}e_x = \frac{3x^2 + 2x + 1}{x^2 + x + 1}r_x$$

Exercise 2

Exercise 2

Estimate the absolute errors in the images of x, with absolute error e_x, under the functions

(i) $x \longmapsto x^3 - 4x + 3 \quad (x \in R)$ (2 minutes)

(ii) $x \longmapsto \dfrac{5}{x} - \dfrac{3}{x^2} \quad (x \in R, x \neq 0)$ (2 minutes)

■

We summarize below the main rules we have obtained in this section for the propagation of errors in evaluating images under functions of one variable of the type we considered.

Summary * * *

Combination of functions of one variable		
Operation	Error estimate	
Addition (or Subtraction)	Add (or Subtract) *Absolute* errors	Rule 5
Multiplication by exact number	Multiply *Absolute** error by exact number	Rule 1
Multiplication (or Division)	Add (or Subtract) *Relative* errors	Rule 3, 4

* The relative error is unchanged in this case.

Discussion

The rules we have just given apply to the estimated errors themselves, not to the error bounds. It is possible to formulate rules for combining estimated error bounds, but we shall not do it here because they are more complicated than the ones for the errors. To get an idea of what can happen you may like to consider the examples below. (*If you have found the work difficult so far, or are short of time, it might be better to skip to the beginning of the next section instead.*)

As a first example, let us obtain an absolute error bound for $x \longmapsto -x$, $(x \in R)$. By the first rule above, the absolute errors in x and $-x$ are related by

$$e_{-x} = -e_x$$

showing that the error in $-x$ has the opposite sign to that of x but the same magnitude. Since only the magnitude of the error affects the error bound, it follows that the error bound for $-x$ is the same as for x:

$$\varepsilon_{-x} = \varepsilon_x$$

As a second example, we consider the estimated absolute error bound for $x \longmapsto x^2$, $(x \in R)$. On page 14 we derived the following estimate for the absolute error:

$$e_{x^2} \simeq 2xe_x$$

Thus the absolute error for x^2 is approximately $2x$ times that for x. Since only the magnitude of $2x$ affects the error bound, it follows that the estimated error bounds are related by

$$\varepsilon_{x^2} \simeq 2|x|\varepsilon_x$$
$$= \begin{cases} 2x\varepsilon_x \text{ if } x > 0 \\ -2x\varepsilon_x \text{ if } x < 0 \end{cases}$$

Notation 1

(the sign ">" means "is greater than" and "<" means "is less than"). This is considerably more complicated than the result $e_{x^2} \simeq 2xe_x$ for the error estimates themselves.

As a last example, to show the effect of addition, say, we calculate an error bound for $x \longmapsto x^2 + x$, $(x \in R)$. From the results just derived, the error estimate is given by the following calculation:

$$e_{x^2} \simeq 2xe_x$$
$$e_x = e_x$$

so that

$$e_{x^2+x} \simeq (2x + 1)e_x$$

by Rule 5.

Since only the magnitude of $2x + 1$ affects the estimated error bound, we have

$$\varepsilon_{x^2+x} \simeq |2x + 1|\varepsilon_x$$
$$= \begin{cases} (2x + 1)\varepsilon_x & \text{if } x > -\frac{1}{2} \\ -(2x + 1)\varepsilon_x & \text{if } x < -\frac{1}{2} \end{cases}$$

since $2x + 1$ is positive if $x > -\frac{1}{2}$ and negative if $x < -\frac{1}{2}$.

One might expect to be able to obtain this result by applying an addition rule to the error bounds for the individual terms x^2 and x, but it is not so. In fact, the results depend in a rather complicated way on whether x

(*continued on page 18*)

Solution 2

(i) Error in the image $= e_{x^3} - 4e_x + 0$. Therefore, estimated error in the image $= (3x^2 - 4)e_x$. (Notice that we have used the fact that $e_{-4x} = -4e_x$.)

(ii) We have already found the error estimates

$$e_{1/x} \simeq -\frac{e_x}{x^2}, \qquad e_{1/x^2} \simeq -\frac{2e_x}{x^3}$$

so that the error estimate in

$$\left(\frac{5}{x} - \frac{3}{x^2}\right) = -\frac{5e_x}{x^2} + \frac{6e_x}{x^3} = \left(\frac{6 - 5x}{x^3}\right)e_x$$

■

(*continued from page 17*)

is less than $-\frac{1}{2}$, greater than 0, or in the range between $-\frac{1}{2}$ and 0, as shown in the following table:

		$x < -\frac{1}{2}$	$-\frac{1}{2} < x < 0$	$0 < x$
ε_{x^2} ε_x	$\simeq$ $=$	$-2x\varepsilon_x$ ε_x	$-2x\varepsilon_x$ ε_x	$2x\varepsilon_x$ ε_x
ε_{x^2+x}	$\simeq$	$-(2x + 1)\varepsilon_x$ $= \varepsilon_{x^2} - \varepsilon_x$	$(2x + 1)\varepsilon_x$ $= \varepsilon_x - \varepsilon_{x^2}$	$(2x + 1)\varepsilon_x$ $= \varepsilon_{x^2} + \varepsilon_x$

Thus for some values of x the estimated error bound in the sum $x^2 + x$ is the sum of those for the individual terms x^2 and x, as we would expect from the rule given earlier for the error in a sum; but for other values of x, the error bound for the sum is the *difference* of the error bounds for the individual terms.

In fact, one could simplify and say that all three estimated error bounds in the table are less than or equal to

$$\varepsilon_{x^2} + \varepsilon_x$$

Since we only have estimates in any case, why not simplify? The answer is, we can simplify, but in so doing we lose some of the better estimates. For instance, consider

$$x \longmapsto x^2 + x \quad (x \in R)$$

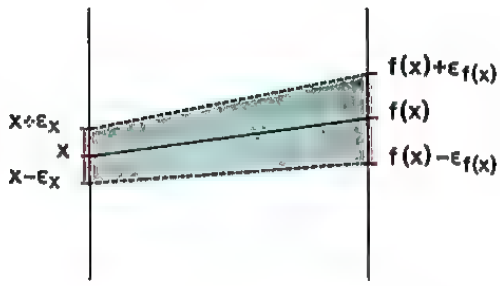
and the image of -2 with absolute error bound of 0.1. From the table (using the first column) we get an estimated error bound of 0.3 for the image; whereas, using our suggested simplification, we get an estimated error bound of 0.5. Nevertheless, in numerical calculations, it is often convenient to use the possibly crude, but simplified, form for the estimated error bound.

2.2.2 Error Intervals

2.2.2
Discussion

In the last section we discovered methods of estimating the error in the image of a number in the domain of a function when we know the error in that number. Generally, of course, we do not have this information; we know only that the number lies in some *interval* in the domain. Can we map this *error interval* in the domain into some *error interval* in the codomain? We can, and we shall find that the ideas we develop will be of use in a branch of mathematics, the solution of equations, which now seems far removed from the present topic but which is considered later in this text and in the television programme.

Consider the mapping diagram shown.



The *error interval* in the domain is known and in this particular case it is determined by the two numbers, x (approximate number) and ε_x (absolute error bound), which are known. The number x maps to the image $f(x)$ in the codomain. We can find the exact images of $x + \varepsilon_x$ and $x - \varepsilon_x$, but usually it is simpler and quicker to estimate these images by the methods of the previous section, and hence find the estimate of the error interval in the codomain from them. The dashed lines in the diagram are meant to indicate the way the *interval* maps under the function and not the images of $x + \varepsilon_x$ and $x - \varepsilon_x$. These upper and lower bounds do not necessarily map respectively to the upper and lower bounds of the image error interval as we see in the next example.

Example 1

Example 1

(i) Determine the error interval in the codomain which corresponds to the error interval [0.4, 0.6] in the domain under the mappings:

(a) $x \longmapsto x^2 \qquad (x \in R)$,

(b) $x \longmapsto \dfrac{1}{1+x} \qquad (x \in R^+)$

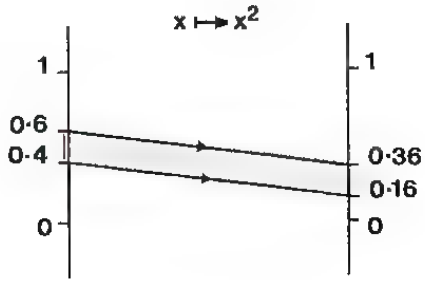
(ii) To what does the error interval $[-0.2, 0.2]$ map under the function (a)?

■

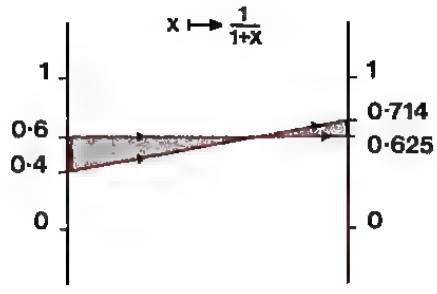
Solution of Example 1

In these cases we can determine the actual bounds directly and need not estimate.

(i) (a) [0.16, 0.36]



(b) [0.625, 0.714]



Notice the crossover. The numbers given in the codomain are accurate to three places of decimals.

If we specified the interval by giving its mid-point 0.5 with absolute error bound 0.1 and used the estimating procedure from the preceding section, we would get

$$\text{absolute error bound in } (1 + x) = 0.1$$

$$\text{relative error bound in } (1 + x) = \frac{0.1}{1 + x}$$

By the division rule, we have

the estimated relative error bound in

$$\frac{1}{1 + x} = \left| -\frac{0.1}{(1 + x)} \right|$$

so that the estimated absolute error bound in

$$\frac{1}{1 + x} = \left| -\frac{0.1}{(1 + x)^2} \right|$$

This holds for all error intervals of half-width 0.1, but we are interested in the one which is centred on 0.5.

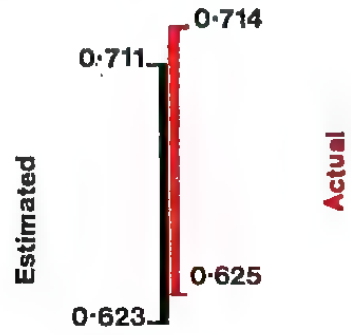
$$\text{Here, estimated absolute error bound} = \frac{0.1}{(1.5)^2} = 0.044$$

$$\text{The image of 0.5 is} \qquad \frac{1}{1.5} = 0.667$$

Thus the estimated error interval in the codomain is

$$[0.667 - 0.044,\ 0.667 + 0.044] = [0.623, 0.711]$$

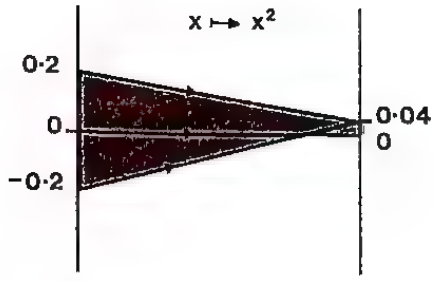
with end-points differing by only 0.002 or 0.003 from the exact ones calculated above.



For *particular* error intervals in the domain the estimation method again took longer. Its power lies in its generality as we shall see in the next exercise.

First we must complete the solution of the example.

(ii)



If we adopt the method of estimation given in (i), the interval appears to shrink to nothing: but consider *some other numbers* in the interval. You will see that the lower end-point in the codomain is the image of zero; so beware. If you use this method, always make a quick check of the numbers inside the interval to make sure that their images are behaving themselves. ■

For the next exercise, we need the following definition.

Definition 1

The scale factor for a function of one variable, propagating an error from the domain to the codomain, is defined as

$$\frac{\text{estimated error in image of } x}{\text{error in } x} \qquad (x \in \text{domain})$$

Note that this definition of the scale factor also gives us an estimate of the ratio

$$\frac{\text{error interval width in codomain}}{\text{error interval width in domain}}$$

since we simply choose two elements, the upper and lower bounds of the interval, in the definition. In other words, if the scale factor is greater than 1 the interval length is magnified, but if it is less than one, the length shrinks. If its sign is negative it implies "crossover", as in the figure in the solution of Example 1(i)(b).

Exercise 1

Exercise 1
(5 minutes)
(See Exercise 2.2.1.2)

Using the results of Exercise 2.2.1.2, calculate the scale factor for the following functions:

(i) $x \longmapsto x^3 - 4x + 3 \quad (x \in R)$

(ii) $x \longmapsto \frac{5}{x} - \frac{3}{x^2} \quad (x \in R, x \neq 0)$

(iii) $x \longmapsto \frac{1}{5}(x^3 + 3) \quad (x \in R)$

at $x = -2, x = \frac{2}{3}, x = 2.$ ■

(This particular exercise will help you to appreciate the television component of this unit.)

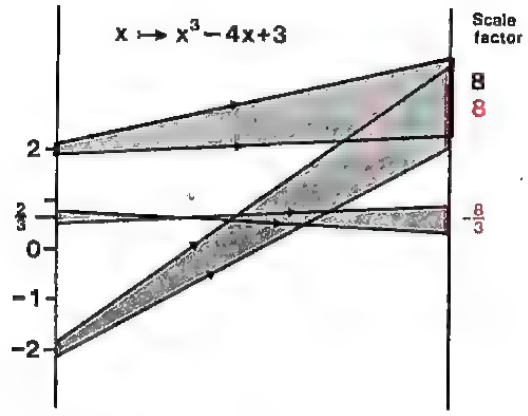
Solution 1

Solution 1

(i) We found in Exercise 2.2.1.2(i) that the estimated error in the image of $(x + e_x)$ was $(3x^2 - 4)e_x$.

Therefore the scale factor is

$$\frac{(3x^2 - 4)e_x}{e_x} = 3x^2 - 4$$

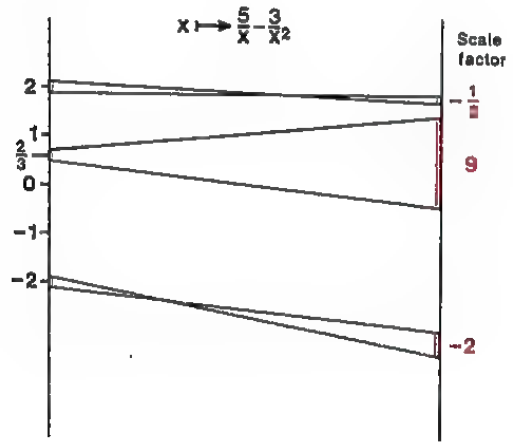


x	scale factor
$+2$	8
$\frac{2}{3}$	$-\frac{8}{3}$
-2	8

The diagram shows how the error intervals in the domain are magnified in the codomain.

(ii) Using the result from Exercise 2.2.1.2(ii), the scale factor is

$$\frac{6 - 5x}{x^3} \quad (x \neq 0)$$



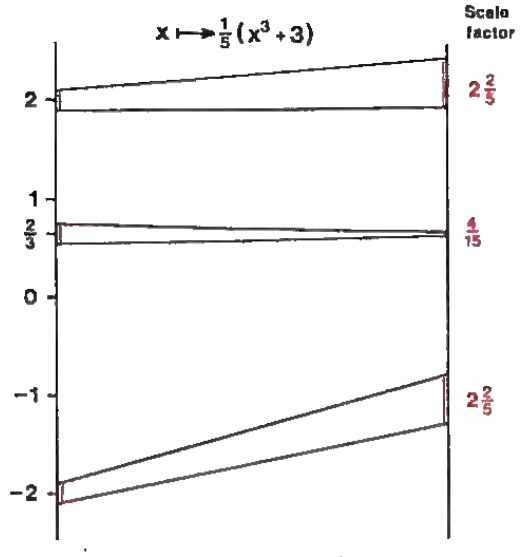
x	scale factor
$+2$	$-\frac{1}{2}$
$\frac{2}{3}$	9
-2	-2

Here, the two error intervals centred on $\frac{2}{3}$ and -2 in the domain grow, but the one centred on 2 shrinks.

(iii) In this case the scale factor is

Solution 1 (*continued*)

$$\frac{3x^2}{5}$$



x	scale factor
$+2$	$\frac{12}{5}$
$\frac{2}{3}$	$\frac{4}{15}$
-2	$\frac{12}{5}$

Here, the two error intervals centred on 2 and -2 in the domain grow, but the one centred on $\frac{2}{3}$ shrinks. No crossover occurs, since all three scale factors are positive. ■

2.3 ERROR PROPAGATION USING FUNCTIONS OF TWO VARIABLES

2.3

2.3.0 Introduction

2.3.0 Introduction

In this section we shall see that the rules for error propagation which we have already developed for functions of one variable do not require any major modification for functions of more variables. For example, the absolute error in the image of (x, y) under the function

$$(x, y) \longmapsto x + y \quad (x \in R, y \in R)$$

is $e_x + e_y$, corresponding to Rule 2.2.1.5 for functions of one variable.

Exercise 1

Exercise 1 (2 minutes)

A bucket containing water weighs 4 kg. When empty it weighs 1.5 kg, each weight being accurate to ± 0.1 kg. Determine the approximate weight of the water and the relative and absolute error bound in this result. ■

Solution 1

Solution 1

The approximate weight of the water is $4 - 1.5 = 2.5$ kg. The greatest possible weight of water is

$$4.1 - 1.4 = 2.7 \text{ kg}$$

The least possible weight of water is

$$3.9 - 1.6 = 2.3 \text{ kg}$$

Hence the absolute error bound is 0.2 kg, and the relative error bound is $\frac{0.2}{2.5} = 0.08$ (or 8 % accuracy). ■

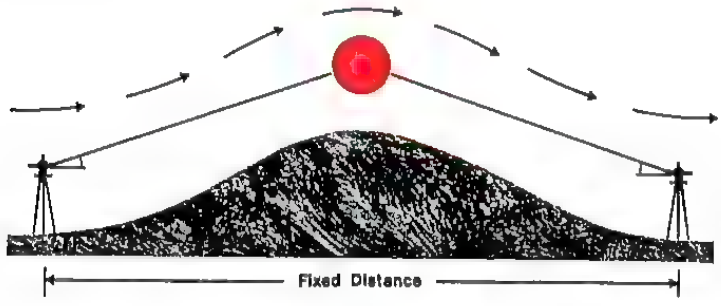
Exercise 2

Exercise 2
(5 minutes)

In an experiment to determine the airflow pattern over a small hill, no-lift balloons (balloons whose weight is balanced exactly by their buoyancy) are observed at five-second intervals, by two theodolites at fixed positions. Suppose that there is a 1 % relative error bound in the heights calculated from the theodolite observations, and that these heights are used to calculate the vertical velocity. In particular, calculate the relative error bound in the vertical velocity

$$\frac{130 - 120}{5} = 2 \text{ m/sec}$$

determined from two consecutive heights of 120 m and 130 m, if the time interval of 5 sec is assumed to be exact. What is the corresponding absolute error bound?



■

The last exercise illustrates the point that when we subtract two nearly equal numbers, the relative error bound increases dramatically, whilst the absolute error bound does not change so much. For example, in the last exercise the relative error bound increases by a factor of 25, whilst the absolute error bound approximately doubles. Since the relative error bound tells us the size of the possible error in relation to the size of the number, we conclude that the result of an operation such as this can be highly suspect, particularly if we use this result in subsequent calculations.

Solution 2

Absolute error bound in 130 m is

$$0.01 \times 130 = 1.3 \text{ m}$$

Absolute error bound in 120 m is

$$0.01 \times 120 = 1.2 \text{ m}$$

In this case, the absolute error bound in the height difference is

$$1.2 + 1.3 = 2.5 \text{ m}$$

and the relative error bound in the height difference is

$$\frac{2.5}{10} = 0.25 \text{ (or } 25\%)$$

When we divide by the time difference of 5 sec, assumed exact, we get an absolute error bound of 0.5 m/sec in the vertical velocity, 2 m/sec, and a relative error bound of 0.25. ■

2.3.1 Multiplication and Division

Exercise 1

Exercise 1
(5 minutes)

You measure the sides of a rectangle to be turfed in your garden as 80 ft and 40 ft, both measurements to within ± 1 ft.

(i) Is it possible to calculate the area to the nearest square foot?
(ii) What is the error interval for the area? ■

Main Text

To find the general rule for error propagation in multiplication we consider the operation of multiplication as a function

$$f:(x, y)\longmapsto xy \quad (x \in R^+, y \in R^+)$$

We learnt in Section 2.2.1 to add relative errors in multiplication in functions of one variable. Let us calculate the relative error of the product xy. The absolute error is

$$\begin{aligned} xy - XY &= xy - (x - e_x)(y - e_y) \\ &= ye_x + xe_y - e_xe_y \end{aligned}$$

and the relative error is therefore

$$r_{xy} = \frac{ye_x + xe_y - e_xe_y}{xy} = \frac{e_x}{x} + \frac{e_y}{y} - \boxed{\text{small } \frac{e_xe_y}{xy}}$$

Ignoring the small term at the end (corresponding to the 1 ft^2 in the last exercise), we obtain an estimated relative error

$$\boxed{r_{xy} \simeq \frac{e_x}{x} + \frac{e_y}{y} = r_x + r_y}$$

corresponding to Rule 2.2.1.3 which we found for relative errors in products of functions of one variable.

Exercise 2

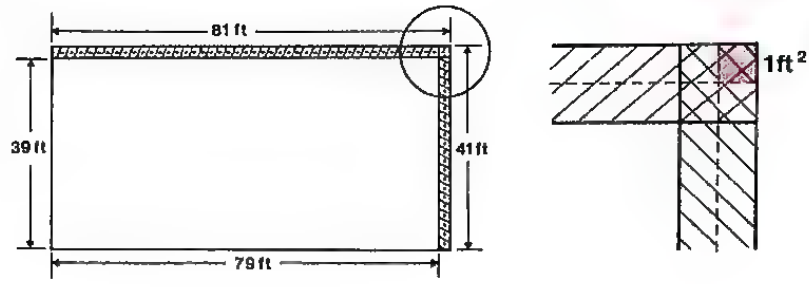
Exercise 2
(2 minutes)

By considering the function

$$f:(x, y)\longmapsto \frac{x}{y} = x \times \frac{1}{y} \quad (x \in R^+, y \in R^+)$$

deduce the rule for propagation of relative errors in division. ■

Solution 1 **Solution 1**



(i) NO. The reason is contained in the solution to part (ii).
(ii) The maximum value of the area (in square feet) is

$$(80 + 1)(40 + 1)$$

$$= \underbrace{80 \times 40}_{\text{approx. area}} + \underbrace{80 \times 1 + 40 \times 1}_{\text{area of two strips along the sides}} + \underbrace{1 \times 1}_{\text{area of small blocked square}}$$

$$= 3200 + 121$$

The minimum value of the area is

$$(80 - 1)(40 - 1)$$

$$= 3200 - 119$$

Thus the error interval is [3081, 3321]. ■

Solution 2 **Solution 2**

Using the rule for multiplication stated just prior to this exercise we get:

$$\text{estimated relative error of } \left(\frac{x}{y}\right) = r_x + r_{1/y} = r_x - r_y$$

the last part follows from Rule 2.2.1.4 on page 15. ■

2.4 ACCURACY IN THE NUMERICAL SOLUTION OF EQUATIONS

2.4

2.4.0 Introduction

2.4.0

Introduction

Frequently the eventual solution to a problem in mathematics depends on the solution of an equation of the general form

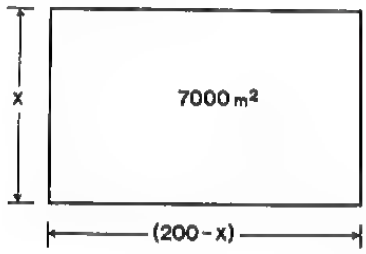
$$f(x) = 0$$

To take a simple example, given that you have to enclose a rectangular pen of area 7000 m^2 with a fence of length 400 m, you may call the length of one side x m, deduce that the other side must have length $(200 - x)$ m and end up with the equation

$$x(200 - x) = 7000$$

to solve. This is equivalent to finding the solution of

$$x^2 - 200x + 7000 = 0$$



Quadratic equations turn up so frequently that probably, in the past, you learnt a general formula for their solution.

A powerful method of solving nearly all equations, including quadratic equations, which adopts a different approach from finding general formulas, is known as the iterative method. In this we make a guess at the solution of the equation and refine it step by step to the desired accuracy. The power of the iterative method lies in its applicability to a wide variety of equation types, in particular to those for which there is no "formula" solution.

Definition 1

2.4.1 Solving a Cubic Equation

2.4.1

Discussion

This section is essentially a summary of part of the television programme associated with this unit. If you have seen, or intend to see, the programme, you need only read this section as a reminder or reinforcement of the points we made there. If you have not seen the programme, this section will help you to get the main idea that it contained, although it should in no way be considered as an adequate substitute.

We consider the problem of solving the cubic equation

$$x^3 - 5x + 3 = 0$$

which has no simple "formula" solution, unlike the quadratic equation we mentioned in the introduction, 2.4.0. One simple way of approximately solving equations of this type is to draw graphs.

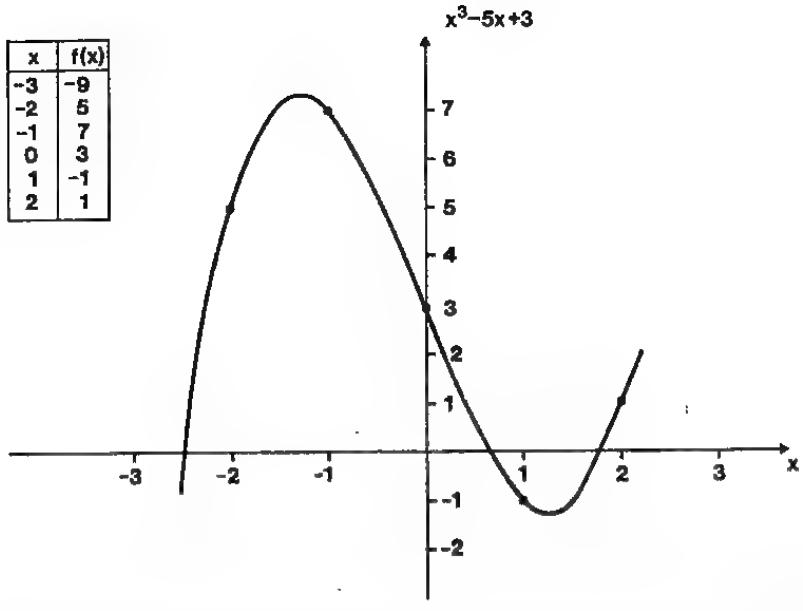
The Graphical Approach

We draw the graph of the function

$$f: x \longmapsto x^3 - 5x + 3 \quad (x \in R)$$

by plotting points derived from the table below and drawing a smooth curve through them.

x	f(x)
-3	-9
-2	5
-1	7
0	3
1	-1
2	1



The values of x where the graph of f crosses the axis will be the solutions of the equation, since they satisfy

$$f(x) = 0$$

that is,

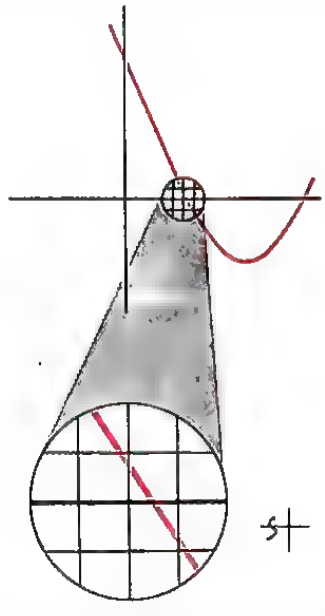
$$x^3 - 5x + 3 = 0$$

From the graph we see that the solutions are in the intervals

$$[-3, -2], \quad [0, 1], \quad [1, 2]$$

and we might well make a guess at the value of the second solution, say as 0.5.

Clearly we could improve the accuracy of this second solution by re-plotting the particular portion of the curve for the interval [0.4, 0.7] on a magnified scale at intervals of 0.1 and then drawing a smooth curve through these points and finding where it cuts the x-axis.



Repetitive steps of this type could achieve any accuracy we require but at great expense in time.

An Iterative Approach

Main Text

We may rearrange

$$x^3 - 5x + 3 = 0$$

into the form

$$x^3 + 3 = 5x$$

and thus into the form

$$x = \tfrac{1}{5}(x^3 + 3)$$

The significant feature we would like you to note is the following one.

IF we can find a value of x which makes $\frac{1}{5}(x^3 + 3)$ equal to x, then that is a solution of our cubic equation; for it makes

$$x^3 - 5x + 3 = 0$$

Let us try an experiment and build up the table below starting with an initial guess 0.5 for x, then calculating the corresponding value of $\frac{1}{5}(x^3 + 3)$ and using this value as our new guess for x and so on.

x	$\frac{1}{5}(x^3+3)$
0·5	0·625
0·625	0·648
0·648	0·654
0·654	0·656
0·656	0·656

The value 0.656 is thus a solution to three places of decimals. In fact we could get any required degree of accuracy. This process is called an *iterative process* and in this case was successful in finding a solution for us.

Let us try some more such experiments. These are tabulated below, and show two attempts to find the solutions in the intervals $[-3, -2]$, $[1, 2]$ respectively using the rearrangement

$$x = \tfrac{1}{5}(x^3 + 3)$$

and three attempts to find solutions using the rearrangement

$$x = x^3 - 4x + 3$$

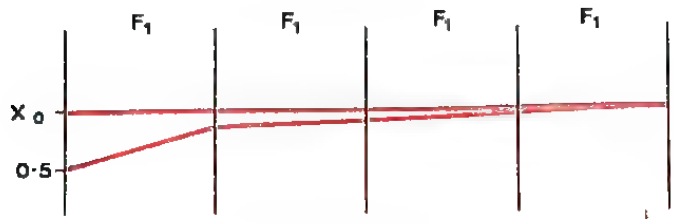
1	2	3	4	5
$x = \frac{1}{5}(x^3 + 3)$	$x = \frac{1}{5}(x^3 + 3)$	$x = x^3 - 4x + 3$	$x = x^3 - 4x + 3$	$x = x^3 - 4x + 3$
$x = -2$	$x = 2$	$x = -2$	$x = 0.5$	$x = 2$
−1.000	2.20	3	1.125	3
0.400	2.73	18	−0.077	18
0.613	4.67	255	2.692	255
0.646	21.0	⋮	11.75	⋮
0.654	1853	⋮	⋮	⋮
0.656	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮

The first guess in each case is shown at the head of the column under the rearrangement used, and the later approximations are then listed in the column below. None of the attempts was successful, except that in 1 we found our solution 0.656 again.

It would be useful to have some criterion enabling us to decide, without too much calculation, whether an iterative method is likely to succeed or fail in any given case. A way of doing this is to consider the function associated with the rearrangement, e.g.

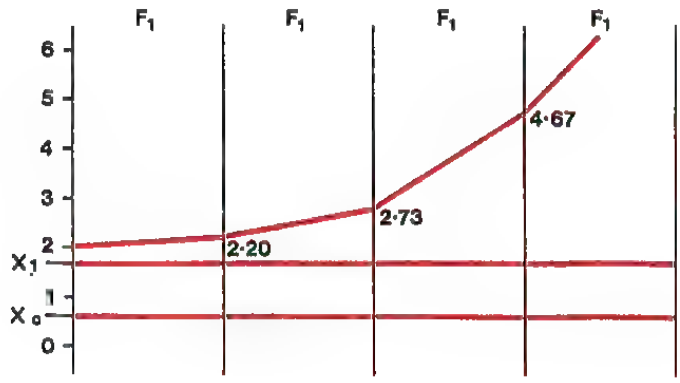
$$F_1 : x \longmapsto \tfrac{1}{5}(x^3 + 3)$$

for the first one. The iterative process can be regarded as an application of a composite function, i.e. as a repeated application of the same function, as indicated in the mapping diagram below.



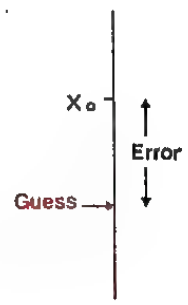
The solution, X_0 say, remains unchanged under F_1 and is thus represented by a horizontal line. Our first guess, 0.5, was improved at each application of the function, and we obtained better and better approximations to the solution.

Starting with $x = 2$ however, in an attempt to find X_1 (the solution in [2, 3]), met with failure, as is illustrated in the mapping diagram below.



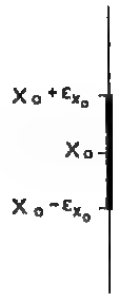
How can we determine the behaviour of our guess in advance? If we look at the domain of the function near the solution X_0 we can interpret our guess as

$$X_0 + \text{error}$$



We would like the function to diminish the error. We can determine whether it does so, by looking at the behaviour of error intervals containing X_0. If they shrink under the function then ANY guess which

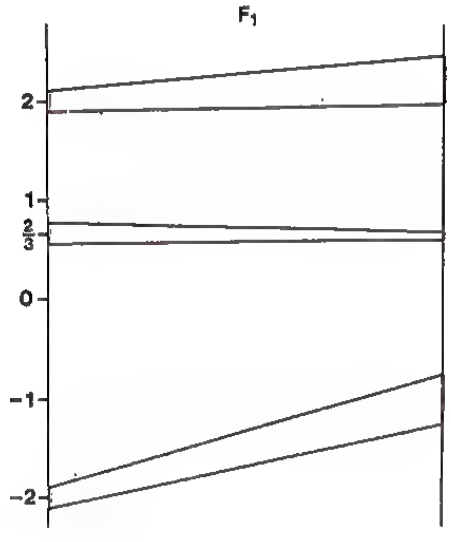
originally lies within one of them is likely to be improved. If they grow, and look as if they are going to continue to grow, we would expect failure. This growth, or shrinkage rate, is measured by the scale factor.



As we saw in the solution to Exercise 2.2.2.1(iii) the scale factors for the error intervals for F_1, near the three solutions, are

x	Scale factor
2	$\frac{12}{5}$
$\frac{2}{3}$	$\frac{4}{15}$
-2	$\frac{12}{5}$

which gave the mapping diagram:



These scale factors thus enable us to predict when the iterative method will work.

Exercise 1

Exercise 1
(5 minutes)

Using Exercise 2.2.2.1(i) and (ii), test which of the solutions you can find by using the rearrangements

(i) $x = x^3 - 4x + 3$

(ii) $x = \dfrac{5}{x} - \dfrac{3}{x^2}$ ■

Solution 1

Solution 1

Copying down the scale factors from Exercise 2.2.2.1, we have

$x \longmapsto x^3 - 4x + 3$

x	Scale factor
2	8
$\frac{2}{3}$	$-\frac{8}{3}$
-2	8

$x \longmapsto \frac{5}{x} - \frac{3}{x^2}$

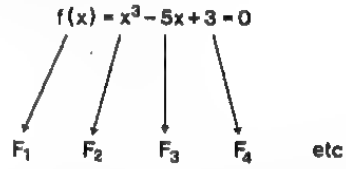
x	Scale factor
2	$-\frac{1}{2}$
$\frac{2}{3}$	9
-2	-2

Clearly the answers are

(i) None. (ii) The solution near 2. ■

There are many rearrangements of our cubic,

Discussion



and we could test any of them using our "error interval" approach. It would be even more useful, however, if we had a method that took us straight to an effective iterating function F without any trial and error. There is such a method. It is called the Newton–Raphson method. The iterative process that it gives for our cubic is the one associated with the function

$$F: x \longmapsto \frac{2x^3 - 3}{3x^2 - 5} \quad (x \in R, x^2 \neq \tfrac{5}{3})$$

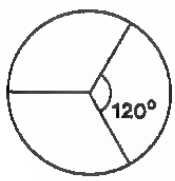
This process will find all three solutions provided our first guesses are reasonably good (e.g. 0.5, 2, −2).

2.4.2 The "Omelette" Problem

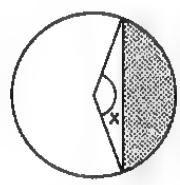
2.4.2

Discussion

If we wish to share a cake equally between 3 persons, we know that theoretically the angle at the centre should be 120°. Omelettes however,



Cake



Omelette

tend to be cut with one straight cut in the frying pan. What, we might theoretically ask, is the angle subtended at the centre for the shaded area to be $\frac{1}{3}$ of the total? With some knowledge of geometry we can deduce that the angle x, measured in radians, is the solution of the equation

$$x - \sin x - \frac{2\pi}{3} = 0$$

What we intend to do in this section is to ask you to find a crude graphical solution to this equation using the method introduced in the first part of Section 2.4.1. Then we will look at a different type of iterative approach from the rearrangement one; the latter cannot be applied so readily in this case, since we do not know how to estimate the scale factors for error intervals under the function $x \longmapsto \sin x$, $(x \in R)$.

Exercise 1

Exercise 1
(10 minutes)

Find the images of $0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi$, to an accuracy of two decimal places, under the mapping

$$f: x \longmapsto x - \sin x - \frac{2\pi}{3} \quad (x \in R);$$

plot the points $(x, f(x))$ for

$$x = 0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi$$

and hence sketch the graph of the function in the interval $[0, \pi]$. Estimate where this crosses the x-axis. ■

We now derive an approximate solution by an iterative method. To two decimal places in Exercise 1 we found that

$$f\left(\frac{2\pi}{3}\right) \simeq f(2.09) \simeq -0.87$$

$$f(\pi) \simeq f(3.14) \simeq 1.05$$

Since one image is positive and one is negative, it implies (as we saw in the graph) that the solution lies somewhere between the two values 2.09 and 3.14, i.e. it lies in the interval [2.09, 3.14].

Exercise 2

Exercise 2
(3 minutes)

If we investigate the sign of the image of the mid-point of the interval [2.09, 3.14], i.e. 2.615, we find that this is positive.

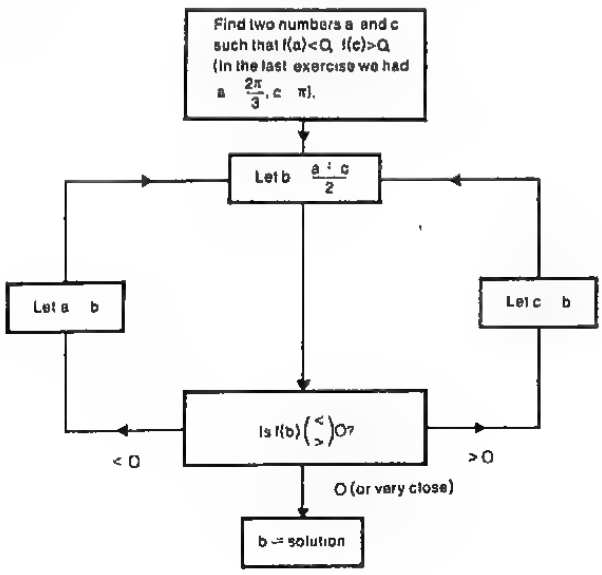
(i) Does the solution lie in the interval [2.09, 2.615] or [2.615, 3.14]?
(ii) What is the ratio of the width of this new "error interval" to that of the old one?
(iii) How could you still further reduce the width of the interval in which the solution is contained? ■

Let us now try to write a general prescription so that you, or a fellow student, could apply it (in theory at least) to find a solution of any equation

Discussion

$$f(x) = 0$$

Try to think about this logically yourself before reading the prescription which follows, which is expressed in the form of a flow diagram, i.e. a schematic form of the steps in the solution. Notice that the symbols a, b, c change their numerical values at various stages. Thus "let $c = b$" means "change the value of c to b". This is different from the usual conventions of algebra, but it is a very useful convention in computer programming.



(continued on page 36)

Solution 1

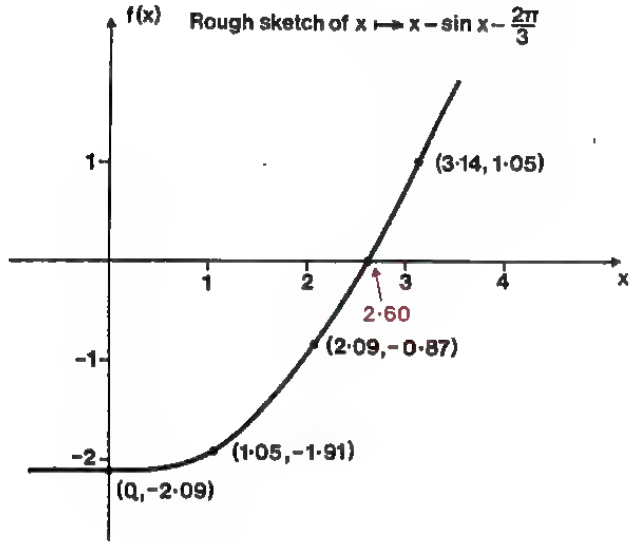
Solution 1

A rough sketch of the curve is drawn below.

Your estimate of the solution of

$$x - \sin x - \frac{2\pi}{3} = 0$$

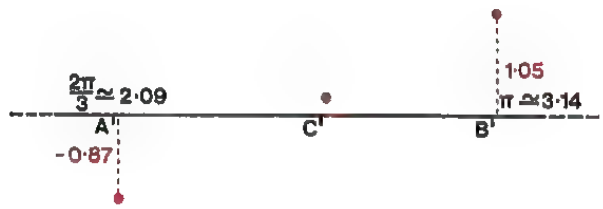
should have been somewhere near 2.6 radians.



■

Solution 2

Solution 2



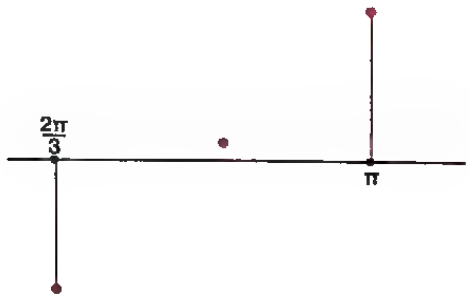
(i) The image at C, the mid-point of AB, is positive. Thus the point representing this is above the axis. The solution lies in the interval $[2.09, 2.615]$.

(ii) $\frac{1}{2}$.

(iii) Find the image of the mid-point of AC and test whether it is positive or negative. This again halves the interval in which the solution is known to lie. ■

(continued from page 35)

We could in fact have determined the number of steps required in advance to achieve a given accuracy by obtaining a general formula for the absolute error bound after n approximations.



When we first discovered that the solution of the omelette problem was in the interval $\left(\frac{2\pi}{3}, \pi\right)$, we knew that the mid-point approximation

$$\frac{\left(\frac{2\pi}{3} + \pi\right)}{2}$$

could not be in error by more than half the width of the interval; so that the error bound of this approximation is

$$\frac{\left(\pi - \frac{2\pi}{3}\right)}{2} = \frac{\pi}{6}$$

For each successive approximation the absolute error bound is reduced by a factor of 2. Thus the next approximation would have an absolute error bound $\frac{\pi}{12}$, and the nth approximation $\frac{\pi}{3 \times 2^n}$.

The construction of a flow diagram, such as the one on page 35, is a valuable tool in programming a computer to do a calculation. In the present case, the calculation was tried out; the computer was told to stop when the width of the error interval was less than $\frac{\pi}{18\,000}$. This was the so-called "zero" in the flow diagram, and it produced the result 2.6053 radians. This corresponds to an angle of about 150° and means that we cut an omelette about $\frac{3}{8}$ of the distance along a diameter from the edge (if it isn't cold by now!). For 12 steps we would have an absolute error bound of $\frac{\pi}{(3 \times 2^{12})} = \frac{\pi}{12\,288}$ and we needed to go to 13 steps $\left(\text{absolute error bound} = \frac{\pi}{24\,576}\right)$ to achieve the desired accuracy. (Note that the accuracy of this method is restricted by the accuracy with which π and $\sin x$ can be found.)

Summary

(i) Graphical methods can nearly always be used to solve equations, given sufficient time.
(ii) Iterative methods are very versatile methods of solving equations.
(iii) Error arithmetic can be used with certain functions to predict whether iteration based on a given rearrangement of the equation will give a desired solution or not.
(iv) In an iterative method we have direct control of the accuracy throughout the computation. In particular, we can improve the accuracy as we proceed through the computation (if the method is known to be a successful one), starting with a rough estimate and refining.
(v) Slips in an iterative calculation are often self-compensating. This last point has not been mentioned before but can be important. A slip might lengthen the *time* to perform the calculation, but otherwise it has little effect unless it is repeated every time.

2.5 BLUNDERS AND THEIR CONTROL

2.5

Discussion

A common source of error in a computation is the blunder. This can take almost any form, but one typical form is the transposition of two digits when copying. For example, 1547 may be copied as 1457. Another one is "1988" for "1998", when the wrong digit is repeated. Measurements too can be read incorrectly and we frequently *check* these by measuring again in a different way, particularly if one of the measurements in a table of results is clearly out of step with the others. We shall see some useful ways of checking for errors in tables of measurements and of mathematical functions in *Unit 4, Finite Differences*. This habit of checking is an important habit to adopt in all numerical work (particularly that involved with the calculation of your salary or wage packet), and is an important defence against blunders. We outline two methods of checking below.

(i) *Check that the size of the result is about right.* In other words, examine its *order of magnitude.*

Example 1

Example 1

Determine the area of a rectangular floor with sides, assumed exact, of 3.6 m and 4.3 m respectively. ■

Solution of Example 1

In multiplying, say we use a slide rule and put the decimal point in wrongly at the end. We may then get a result of 154.8 m^2. A quick check of $(3 \times 4) = 12$ tells us that we have the decimal point in the wrong place. The order-of-magnitude check does not imply that we now have the correct answer. In copying the magnitude of the area on to another piece of paper we may write 15.84 m^2. We might also have made a mistake in reading the slide-rule, or, if the calculation is done with pencil and paper, in one of only too many ways:

```
e.g. 3.6
      4.3
     ----
     10 8
    134      (instead of 144)
   -----
   14.48
```

■

(ii) *Repeat the calculation in a different way.* Simple repetition of the same method often leads to a repetition of the same error (as you have probably found before now), and is the worst method of checking. For example, in the particular calculation we are studying, a good check would be to multiply the other way:

```
     4.3
     3.6
    ----
     25 8
    129
   -----
   15.48
```

The different result shows that there is a blunder somewhere.

Exercise 1

Exercise 1
(5 minutes)

The daily sale of bottles of milk on a milk round during one week is illustrated below. Find the total for the week by your usual method and try to find two other ways of obtaining the same total.

Sunday	1762
Monday	1651
Tuesday	1601
Wednesday	1693
Thursday	1687
Friday	1736
Saturday	1859

■

Three possible methods of forming the sum are outlined below.

Adding *down* columns from right	Adding *up* columns from right	Addition of differences from 1700	
1762	1762	62	
1651	1651		−49
1601	1601		−99
1693	1693		− 7
1687	1687		−13
1736	1736	36	
1859	1859	159	
11989	11989	257	−168 = 89

$7 \times 1700 = 11\,900$

Total $= 11\,989$ ■

2.6 CONCLUSION

In this work you have seen how in mathematics we define the accuracy of an approximation by defining a bound on its error, how we can estimate the magnitude of the errors that propagate through a calculation, and how we can set up defences against blunders.

You have also seen how we could make the error arithmetic that we devised work for us in improving the accuracy of solutions to equations. The two parts are well summed up by the phrase — errors that grow and errors that die — the error arithmetic being useful in the analysis of both.

M100—MATHEMATICS FOUNDATION COURSE UNITS

1 Functions
2 Errors and Accuracy
3 Operations and Morphisms
4 Finite Differences
5 NO TEXT
6 Inequalities
7 Sequences and Limits I
8 Computing I
9 Integration I
10 NO TEXT
11 Logic I—Boolean Algebra
12 Differentiation I
13 Integration II
14 Sequences and Limits II
15 Differentiation II
16 Probability and Statistics I
17 Logic II—Proof
18 Probability and Statistics II
19 Relations
20 Computing II
21 Probability and Statistics III
22 Linear Algebra I
23 Linear Algebra II
24 Differential Equations I
25 NO TEXT
26 Linear Algebra III
27 Complex Numbers I
28 Linear Algebra IV
29 Complex Numbers II
30 Groups I
31 Differential Equations II
32 NO TEXT
33 Groups II
34 Number Systems
35 Topology
36 Mathematical Structures